

Análise exploratória de dados

Sara Mortara, Andrea Sanchez-Tapia, Diogo S. B. Rocha

aula 05

sobre a aula

1. análise exploratória de dados
2. estatísticas descritivas
3. gráficos
4. relações entre variáveis

1. análise exploratória de dados (AED)

a vida sem análise exploratória de dados



conheça seus dados!



objetivos da AED

objetivos da AED

1. controlar a qualidade dos dados

objetivos da AED

1. controlar a qualidade dos dados
2. sugerir hipóteses para os padrões observados

objetivos da AED

1. controlar a qualidade dos dados
2. sugerir hipóteses para os padrões observados
3. apoiar a escolha dos procedimentos estatísticos de testes de hipótese

objetivos da AED

1. controlar a qualidade dos dados
2. sugerir hipóteses para os padrões observados
3. apoiar a escolha dos procedimentos estatísticos de testes de hipótese
4. avaliar se os dados atendem às premissas dos procedimentos estatísticos escolhidos

objetivos da AED

1. controlar a qualidade dos dados
2. sugerir hipóteses para os padrões observados
3. apoiar a escolha dos procedimentos estatísticos de testes de hipótese
4. avaliar se os dados atendem às premissas dos procedimentos estatísticos escolhidos
5. indicar novos estudos e hipóteses

alerta!

alerta!

análise exploratória não é **tortura** de dados



“If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!”

alerta!

análise exploratória não é **tortura** de dados



“If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!”

assume-se que pesquisador(a) formulou *a priori* **hipóteses** plausíveis amparadas pela **teoria**

dicas

dicas

- ▶ pode levar entre 20 e 50% do tempo das análises

dicas

- ▶ pode levar entre 20 e 50% do tempo das análises
- ▶ deve ser iniciada ainda durante a coleta de dados

dicas

- ▶ pode levar entre 20 e 50% do tempo das análises
- ▶ deve ser iniciada ainda durante a coleta de dados
- ▶ utiliza-se largamente técnicas visuais



importância do gráfico e quarteto de Anscombe

- ▶ criado pelo matemático Francis Anscombe
- ▶ 4 conjuntos de dados com as mesmas estatísticas descritivas, mas muito diferentes graficamente



os dados de Anscombe

```
# claro que o conjunto já existe dentro do R  
data("anscombe")
```

```
# média dos dados  
apply(anscombe, 2, mean)
```

```
##          x1          x2          x3          x4          y1          y2          y3          y4  
## 9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909
```

```
# variância dos dados  
apply(anscombe, 2, var)
```

```
##          x1          x2          x3          x4          y1          y2          y3  
## 11.000000 11.000000 11.000000 11.000000 4.127269 4.127629 4.122620  
##          y4  
## 4.123249
```

vamos olhar para os dados

##	x1	x2	x3	x4	y1	y2	y3	y4
## 1	10	10	10	8	8.04	9.14	7.46	6.58
## 2	8	8	8	8	6.95	8.14	6.77	5.76
## 3	13	13	13	8	7.58	8.74	12.74	7.71
## 4	9	9	9	8	8.81	8.77	7.11	8.84
## 5	11	11	11	8	8.33	9.26	7.81	8.47
## 6	14	14	14	8	9.96	8.10	8.84	7.04
## 7	6	6	6	8	7.24	6.13	6.08	5.25
## 8	4	4	4	19	4.26	3.10	5.39	12.50
## 9	12	12	12	8	10.84	9.13	8.15	5.56
## 10	7	7	7	8	4.82	7.26	6.42	7.91
## 11	5	5	5	8	5.68	4.74	5.73	6.89

correlação entre x e y

```
# correlação  
cor(anscombe$x1, anscombe$y1)
```

```
## [1] 0.8164205
```

```
cor(anscombe$x2, anscombe$y2)
```

```
## [1] 0.8162365
```

```
cor(anscombe$x3, anscombe$y3)
```

```
## [1] 0.8162867
```

```
cor(anscombe$x4, anscombe$y4)
```

```
## [1] 0.8165214
```

coeficientes da regressão linear de x e y

```
# coeficientes da regressão
```

```
coef(lm(anscombe$y1 ~ anscombe$x1))
```

```
## (Intercept) anscombe$x1
```

```
##      3.0000909      0.5000909
```

```
coef(lm(anscombe$y2 ~ anscombe$x2))
```

```
## (Intercept) anscombe$x2
```

```
##      3.000909      0.500000
```

```
coef(lm(anscombe$y3 ~ anscombe$x3))
```

```
## (Intercept) anscombe$x3
```

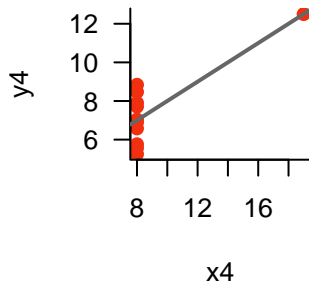
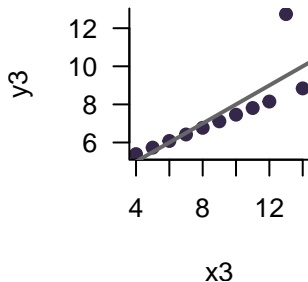
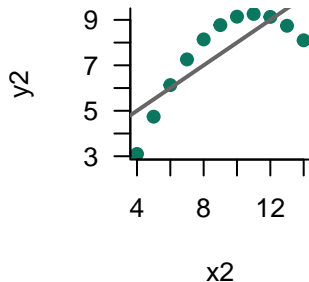
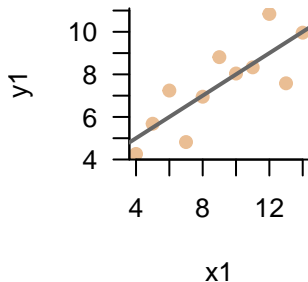
```
##      3.0024545      0.4997273
```

```
coef(lm(anscombe$y4 ~ anscombe$x4))
```

```
## (Intercept) anscombe$x4
```

```
##      3.0017273      0.4999091
```


agora sim vamos olhar para os dados do Anscombe



perguntas que nos devemos fazer

1. Onde os dados estão centrados? Como os dados estão distribuídos? Os dados são simétricos, assimétricos, bimodais?
2. Existem outliers?
3. As variáveis seguem uma distribuição normal?
4. Existem relações entre as variáveis? As relações entre variáveis são lineares?
5. As variáveis precisam ser transformadas?
6. O esforço amostral foi o mesmo para cada observação ou variável?

2. estadísticas descriptivas

conferência de dados no R

```
# lendo os dados da idade da população que usa fraldas  
fraldas <- read.csv("data/idade_fraldas.csv")
```

checando os dados

```
# checando os dados  
head(fraldas)
```

```
##      indivíduo idade  
## 1           1     1  
## 2           2    NA  
## 3           3     2  
## 4           4     0  
## 5           5     1  
## 6           6     0
```

```
tail(fraldas)
```

```
##      indivíduo idade  
## 95           95    77  
## 96           96    79  
## 97           97    87  
## 98           98    85  
## 99           99    91  
## 100          100    86
```

inspecionando os dados

```
str(fraldas)
```

```
## 'data.frame':    100 obs. of  2 variables:  
## $ indivíduo: int  1 2 3 4 5 6 7 8 9 10 ...  
## $ idade    : int  1 NA 2 0 1 0 0 1 0 0 ...
```

```
summary(fraldas)
```

```
##      indivíduo      idade  
## Min.   : 1.00   Min.   : 0.00  
## 1st Qu.: 25.75   1st Qu.: 0.00  
## Median : 50.50   Median : 1.00  
## Mean   : 50.50   Mean    :17.17  
## 3rd Qu.: 75.25   3rd Qu.: 3.00  
## Max.   :100.00   Max.    :99.00  
##                      NA's    :2
```

perguntas que devemos fazer aos dados #1

perguntas que devemos fazer aos dados #1

1. existem valores faltantes i.e. (NAs)? Eles são mesmo faltantes?

onde está NA

```
which(is.na(fraldas$idade))
```

```
## [1]  2 17
```

```
fraldas[c(2,17),]
```

```
##      indivíduo idade
## 2             2    NA
## 17            17    NA
```

vamos substituir NA por 0

```
fraldas$idade[is.na(fraldas$idade)] <- 0
```

conferindo se tem NA

```
is.na(fraldas$idade)
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE
```

```
sum(is.na(fraldas$idade))
```

```
## [1] 0
```

perguntas que devemos fazer aos dados #2

perguntas que devemos fazer aos dados #2

2. existem muitos zeros?

```
fraldas$idade==0
```

```
## [1] FALSE TRUE FALSE TRUE FALSE TRUE TRUE FALSE TRUE TRUE TRUE
## [12] TRUE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE TRUE
## [23] TRUE FALSE TRUE FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE
## [34] TRUE FALSE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE TRUE
## [45] TRUE FALSE TRUE FALSE TRUE FALSE FALSE TRUE FALSE TRUE TRUE
## [56] FALSE FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE
## [67] FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE
## [78] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE
```

quantos?

```
sum(fraldas$idade==0)
```

```
## [1] 36
```

perguntas que devemos fazer aos dados #3 #4 #5

perguntas que devemos fazer aos dados #3 #4 #5

3. onde os dados estão centrados? como estão espalhados? são simétricos? enviesados, bimodais?

perguntas que devemos fazer aos dados #3 #4 #5

3. onde os dados estão centrados? como estão espalhados? são simétricos? enviesados, bimodais?
4. existem valores extremos (outliers)?

perguntas que devemos fazer aos dados #3 #4 #5

3. onde os dados estão centrados? como estão espalhados? são simétricos? enviesados, bimodais?
4. existem valores extremos (outliers)?
5. qual a distribuição da variável?

```
summary(fraldas$idade)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.00	0.00	1.00	16.83	3.00	99.00

medidas de tendência central

```
# media  
mean(fraldas$idade)
```

```
## [1] 16.83
```

```
# mediana  
median(fraldas$idade)
```

```
## [1] 1
```

```
# valor mais frequente na amostra  
freqf <- sort(table(fraldas$idade), decreasing = TRUE)  
freqf[1]
```

```
## 0
```

```
## 36
```

medidas de dispersão

```
# variancia  
var(fraldas$idade)
```

```
## [1] 1046.446
```

```
# desvio padrão  
sd(fraldas$idade)
```

```
## [1] 32.34881
```

```
# coeficiente de variação  
sd(fraldas$idade)/mean(fraldas$idade)*100
```

```
## [1] 192.2092
```

```
# intervalo  
range(fraldas$idade)
```

```
## [1] 0 99
```

```
diff(range(fraldas$idade))
```

```
## [1] 99
```

quantis e intervalo inter-quantil (IIQ)

```
# quantis  
quantile(fraldas$idade)
```

```
##    0%   25%   50%   75%  100%  
##     0     0     1     3    99
```

```
# lembrando da saída do summary  
summary(fraldas$idade)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
##      0.00   0.00   1.00   16.83   3.00   99.00
```

```
# mudando os quantis  
quantile(fraldas$idade, probs=c(0.05, 0.5, 0.95))
```

```
##      5%   50%   95%  
##     0.0   1.0  87.1
```

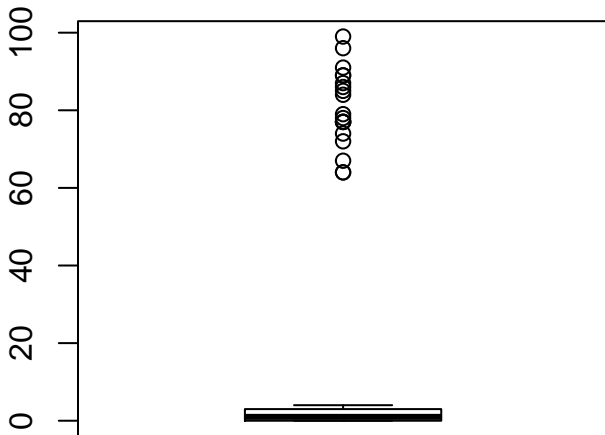
```
# intervalo inter-quantil  
IQR(fraldas$idade)
```

```
## [1] 3
```

3. gráficos

visualizando os dados em um boxplot

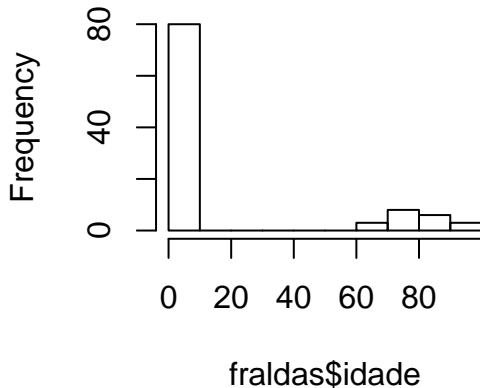
```
boxplot(fraldas$idade)
```



visualizando os dados em um histograma

```
hist(fraldas$idade)
```

Histogram of fraldas\$idade

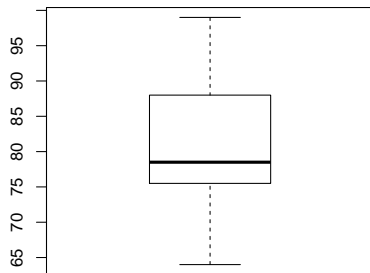
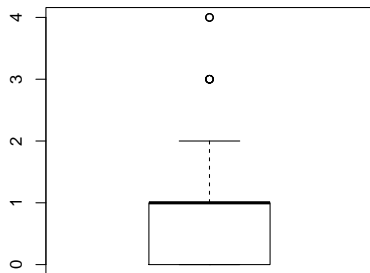


separando bebês e vovxs

```
bb <- fraldas[fraldas$idade<10,]  
vv <- fraldas[fraldas$idade>10,]
```

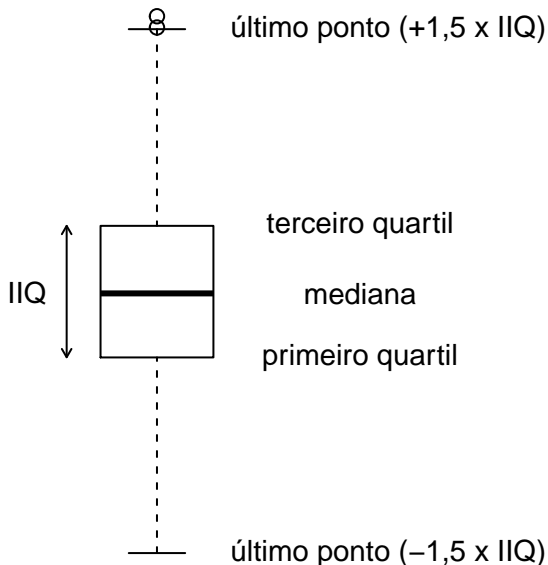

os novos gráficos: boxplot

```
par(mfrow=c(1,2))  
boxplot(bb$idade)  
boxplot(vv$idade)
```



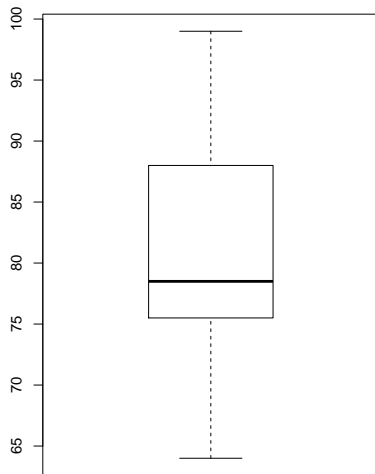
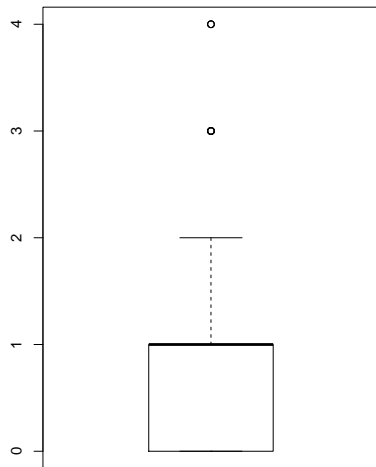
```
par(mfrow=c(1,1))
```

entendendo o boxplot



entendendo o boxplot

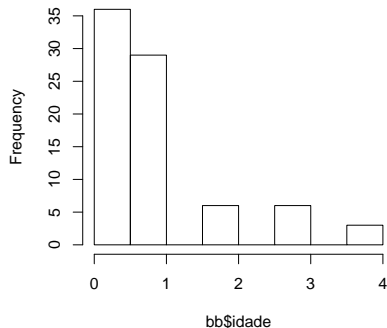
```
par(mfrow=c(1,2))  
boxplot(bb$idade)  
boxplot(vv$idade)
```



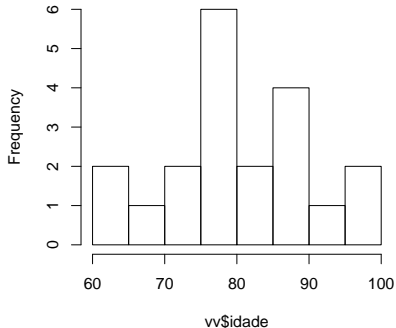
os novos gráficos: histograma

```
par(mfrow=c(1,2))  
hist(bb$idade)  
hist(vv$idade)
```

Histogram of bb\$idade



Histogram of vv\$idade

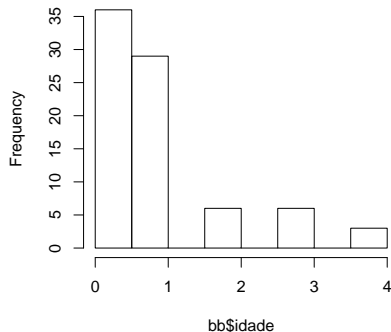


```
par(mfrow=c(1,1))
```

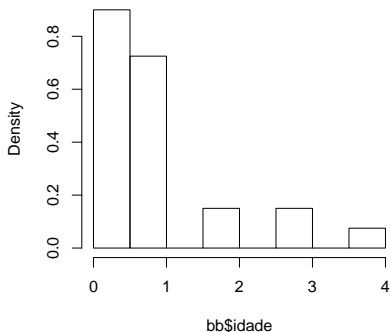
tipos de histograma

```
par(mfrow=c(1,2))  
hist(bb$idade)  
hist(bb$idade, probability = TRUE)
```

Histogram of bb\$idade



Histogram of bb\$idade

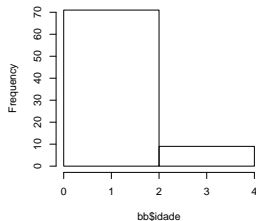


```
par(mfrow=c(1,1))
```

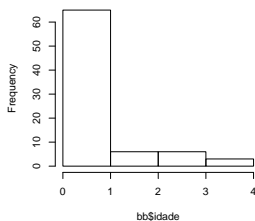
classes do histograma

```
par(mfrow=c(1,3))  
hist(bb$idade, breaks=seq(0, max(bb$idade), length=3))  
hist(bb$idade, breaks=seq(0, max(bb$idade), length=5))  
hist(bb$idade)
```

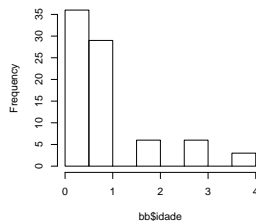
Histogram of bb\$idade



Histogram of bb\$idade



Histogram of bb\$idade



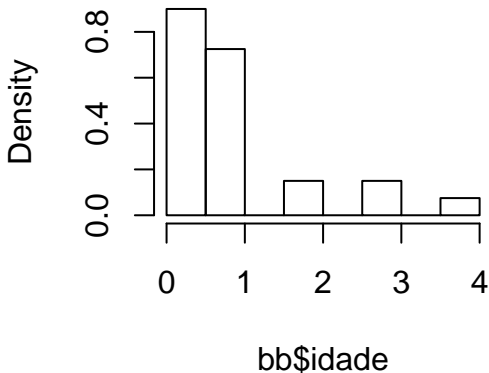
```
par(mfrow=c(1,1))
```

curvas empíricas de densidade probabilística

representa a função que descreve a probabilidade de se encontrar determinado valor

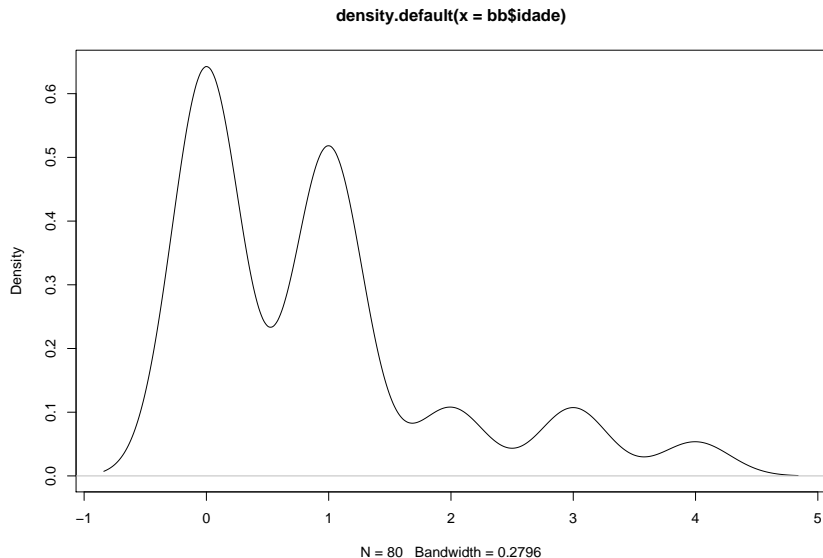
```
hist(bb$idade, probability = TRUE )
```

Histogram of bb\$idade



curvas empíricas de densidade probabilística

```
plot(density(bb$idade))
```



distribuição se ajusta aos dados?

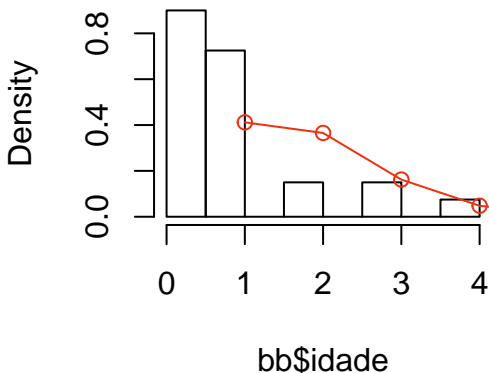
distribuição discreta e assimétrica → Poisson?

```
# máximo de idade  
bb.max <- max(bb$idade)  
# lambda  
bb.med <- mean(bb$idade)
```

distribuição **Poisson** se ajusta aos dados?

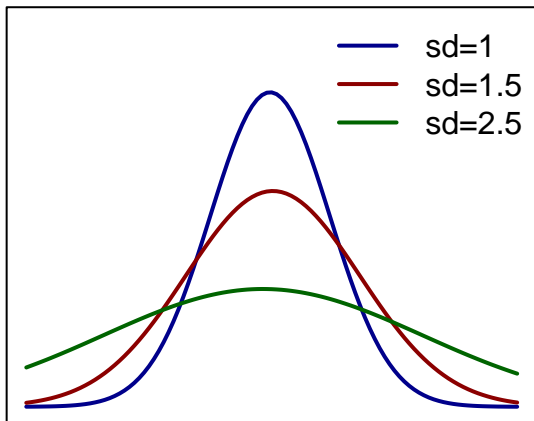
```
hist(bb$idade, probability = TRUE)  
points(dpois(0:bb.max, bb.med), col=cor[5])  
lines(dpois(0:bb.max, bb.med), col=cor[5])
```

Histogram of bb\$idade



(distribuições estatísticas)

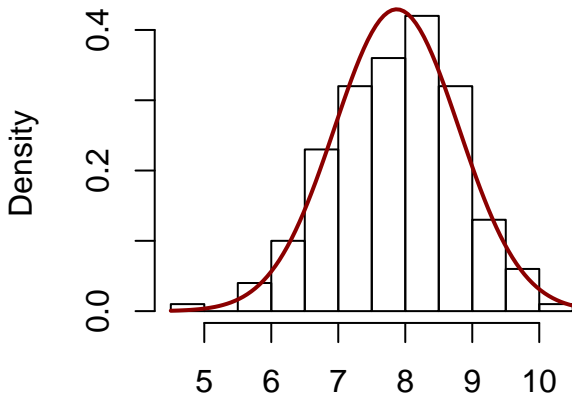
distribuição normal ou gaussiana



por que amostragem é importante?

```
a <- rnorm(200, 8, 1)
hist(a, prob=TRUE)
curve(dnorm(x, mean(a), sd(a)),
      col="darkred", lwd=2, add=TRUE)
```

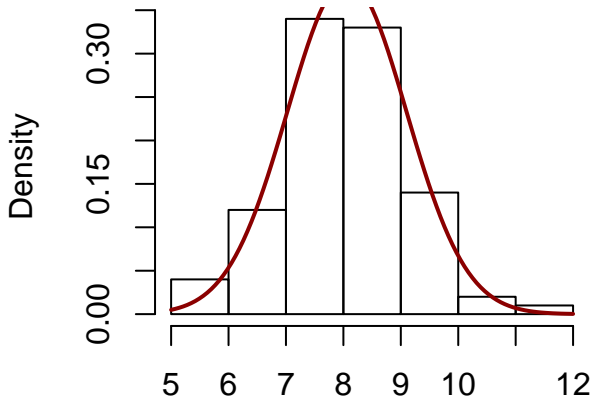
Histogram of a



por que amostragem é importante?

```
a <- rnorm(100, 8, 1)
hist(a, prob=TRUE)
curve(dnorm(x, mean(a), sd(a)),
      col="darkred", lwd=2, add=TRUE)
```

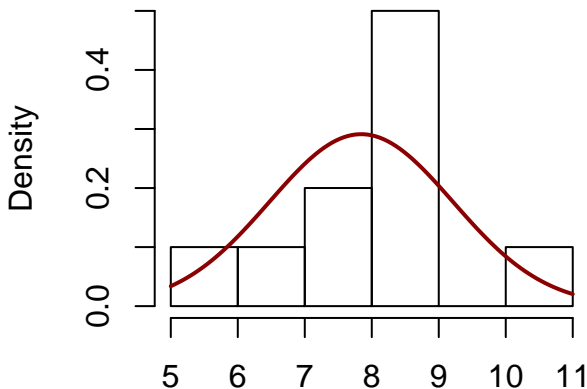
Histogram of a



por que amostragem é importante?

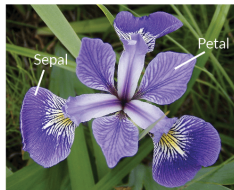
```
a <- rnorm(10, 8, 1)
hist(a, prob=TRUE)
curve(dnorm(x, mean(a), sd(a)),
      col="darkred", lwd=2, add=TRUE)
```

Histogram of a



4. relações entre variáveis

Anderson & Fisher e as espécies de *Iris*



Iris Versicolor



Iris Setosa



Iris Virginica

Anderson & Fisher e as espécies de *Iris*

```
# carregando os dados no R
data(iris)
# para saber mais sobre o conjunto de dados consulte
# ?iris
# entendendo iris
summary(iris)
```

```
##      Sepal.Length      Sepal.Width      Petal.Length      Petal.Width
##  Min.       :4.300      Min.       :2.000      Min.       :1.000      Min.       :0.100
## 1st Qu.:5.100      1st Qu.:2.800      1st Qu.:1.600      1st Qu.:0.300
## Median :5.800      Median :3.000      Median :4.350      Median :1.300
## Mean    :5.843      Mean    :3.057      Mean    :3.758      Mean    :1.199
## 3rd Qu.:6.400      3rd Qu.:3.300      3rd Qu.:5.100      3rd Qu.:1.800
## Max.    :7.900      Max.    :4.400      Max.    :6.900      Max.    :2.500
##           Species
##  setosa      :50
##  versicolor:50
##  virginica   :50
##
##
##
```

correlação entre as variáveis

```
cor(iris[1:4])
```

##	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
## Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
## Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
## Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
## Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

quando uma correlação é alta? 0.7

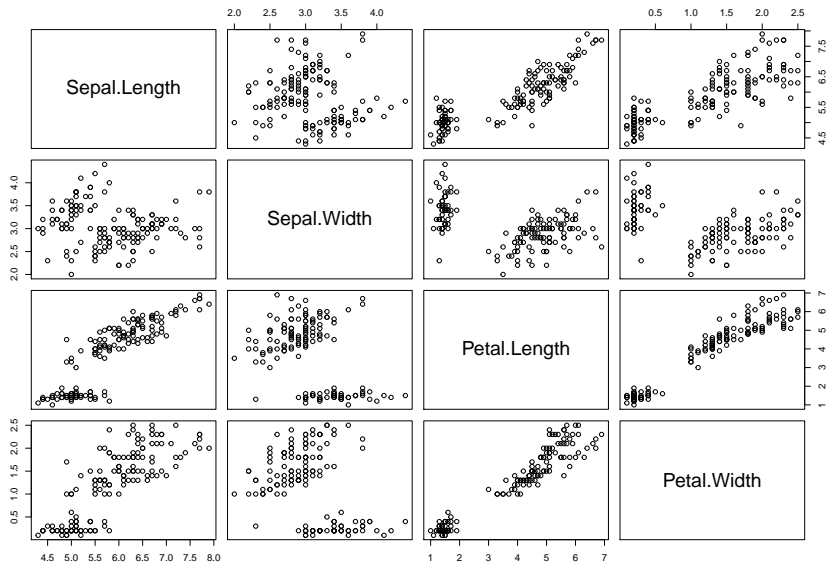
quando uma correlação é alta? 0.7



GENERAL RULE OF THUMB

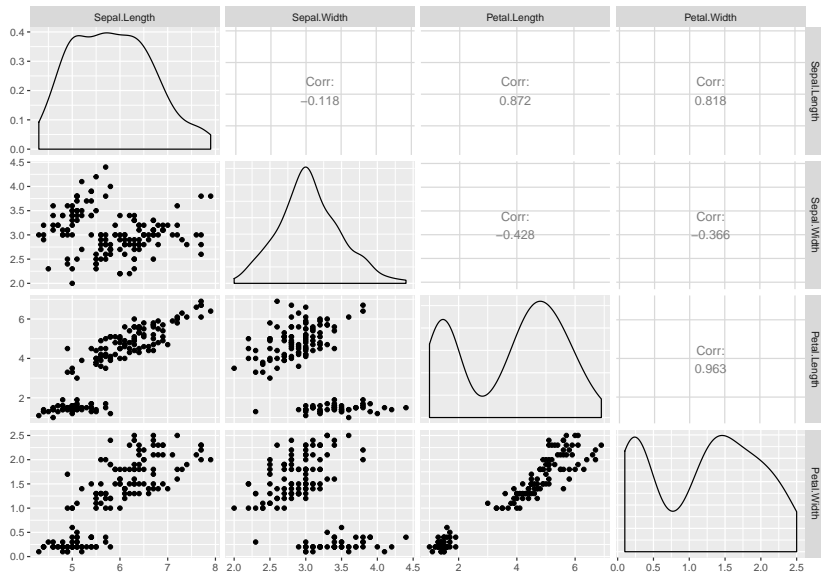
correlação entre as variáveis

```
pairs(iris[1:4])
```



ou ainda melhor correlação entre as variáveis

pacote **Gally** com a função `ggpairs()`



e quais os caminhos para a análise?

sua [H I P Ó T E S E]

depois da [H I P Ó T E S E], quais os caminhos

depois da [H I P Ó T E S E], quais os caminhos

1. entender bem os dados

depois da [**H I P Ó T E S E**], quais os caminhos

1. entender bem os dados
2. variável resposta é normal? → lm e outras análises paramétricas

depois da [**H I P Ó T E S E**], quais os caminhos

1. entender bem os dados
2. variável resposta é normal? → lm e outras análises paramétricas
3. variável resposta tem outra distribuição → análises não paramétricas, glm

depois da [**H I P Ó T E S E**], quais os caminhos

1. entender bem os dados
2. variável resposta é normal? \rightarrow lm e outras análises paramétricas
3. variável resposta tem outra distribuição \rightarrow análises não paramétricas, glm
4. variáveis preditoras hierarquizadas? \rightarrow (g)lmm

depois da [**H I P Ó T E S E**], quais os caminhos

1. entender bem os dados
2. variável resposta é normal? \rightarrow lm e outras análises paramétricas
3. variável resposta tem outra distribuição \rightarrow análises não paramétricas, glm
4. variáveis preditoras hierarquizadas? \rightarrow (g)lmm
5. pseudo-replicação no espaço ou no tempo \rightarrow (g)lmm